Reproducibility of Principal and Independent Genomic Components of Brain Structure and Function

Sourena Soheili-Nezhad[1], Christian F. Beckmann[1], Emma Sprooten*[1]

[1]Departments of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands

*emma.sprooten@donders.ru.nl

Abstract

Introduction

Human brain structure and function, as measured using magnetic resonance imaging (MRI), is heritable. The genetic associations are highly polygenic and pleiotropic, complicating translation of genetic association studies to concrete biological processes. We recently proposed genomic independent component analysis (ICA) to reduce a large array of genome-wide association study (GWAS) statistics into a number of more interpretable multivariate components. Here, we extend this work with genomic principal component analysis (PCA). We also estimate the inter-trait stability and test-retest reliability of genomic PCA and ICA. As a benchmark, we provide a head-to-head comparison with the reproducibility of original (i.e., raw univariate) GWAS variant effect sizes from which the components were derived.

Methods

We considered summary statistics of two rounds of brain MRI GWAS in UK biobank encompassing 22,138 and 11,086 subjects and thousands of imaging-derived phenotypes (IDPs) retrieved from Oxford BIG-40 server. We decomposed the clumped brain-wide genome-wide matrix (n=157,893 variants and m=2,240 brain IDPs) by PCA and ICA to obtain up to 50 principal and independent genome-wide components. We tested the reproducibility of these genomic components using tail overlap coefficient and Pearson's correlation. Decompositions were repeated 20 times on randomly permuted split-half matrices to arrive at an empirical distribution of the above reproducibility metrics. As a benchmark, we also provide comparisons with the reproducibility of the original univariate GWAS effect sizes from which the components were derived.

Results

The first ten principal genomic components were highly stable across trait-subsets within samples (r=0.82±0.15; mean ± standard deviation). PCA cross-sample reproducibility dropped to r=0.16±0.07 (r-max=0.30) was still a substantial improvement over univariate GWAS reproducibility r=0.09±0.05. ICA components showed higher within-sample reproducibility at lower dimensions (ICA dimension-5: r=0.81±0.16 vs. ICA dimension-50: r=0.41±0.13). The same trend was observed for their cross-sample reproducibility, achieving an r-max of 0.27 at the lowest ICA dimension of five (average r=0.20±0.3) vs. r-max of 0.11 at dimension 50 (average r=0.05±0.02).

Conclusion

Overall, the cross-sample reproducibility of PCA and ICA components are moderate, but considerably better than the reproducibility of the raw univariate GWAS effect sizes. Genomic PCA identified highly

reproducible and pleiotropic genomic effects across many brain traits in just 1-3 factors. Genomic ICA captured mutually independent genome-wide effects that may be more sensitive to distinct biological pathways. As GWAS samples continue to increase and more (endo)phenotypes become available in biobanks, multi-trait genome-wide associations of the brain decomposed by ICA and PCA will likely become more reproducible and interpretable.

Introduction

Background

The biological mechanisms that drive individual differences in brain structure and function are complex and largely elusive. Human brain structure and function, as measured using *in vivo* magnetic resonance imaging (MRI), is heritable[1,2], which indicates that genetic research may help to better understand their mechanistic underpinnings. To a large extent the heritability of many human brain features can be captured by common genetic variation as measured by genome-wide association studies[3,4] (GWAS). Similar to other multifactorial traits such as behavioral and psychiatric traits[5], the polygenic nature of brain traits complicates translation of GWAS into an understanding of (causal) biological processes[6]. Analytical algorithms that can help translate high-dimensional MRI-GWAS output into mechanistic interpretations and quantify interindividual biological heterogeneity, could have a tremendous impact both in research, and potentially for clinical applications in personalized medicine.

The richness of big genotypic and phenotypic data in recent biobanks and multi-center consortia makes a new class of post-GWAS analysis possible. For most neurological, psychiatric and behavioral traits, GWAS yields an extremely polygenic signal that is distributed across millions of single nucleotide polymorphisms (SNPs) with very small effect sizes. GWAS of complex brain-related phenotype reflects the far-end outcome of a combination of many different biological pathways from DNA variation to proteins, brain cells and their microenvironment, brain morphology, connectivity and function. In this context, observed genome-wide effects reflect a 'mixture' of many underlying processes, which may be approximated by latent factors. On this basis, multivariate methods that recover the concerted effects of coherent genetic factors across multiple traits may help elucidate the intermediate biological mechanisms. Thus, the inter-dependence of many genetic effects across lots of brain-related phenotypes (i.e. pleiotropy and polygenicity), may hold a key to interpret complex mechanisms underlying patterns of genome-phenome associations. Based on this and using the GWAS of thousands of structural and functional brain MRI traits, we recently proposed genomic independent component analysis (ICA[7]). Here, we extend this proof-of-concept work by applying it to a larger updates GWAS database, also performing genomic principal component analysis (PCA), and assessing inter-trait stability and test-retest reliability in comparison to the raw GWAS data from which the genomic components have been derived.

A few methods for investigating multi-locus genome-wide associations can be considered. One is testing for statistical epistasis, i.e. for interactions between genetic variants. Genome-wide epistasis across millions of variants rapidly becomes computationally intractable in three-way or higher order interactions, is severely penalized by multiple comparisons and has been recently shown to be sensitive to imperfectly tagged causal variants[9]. More efficient are multivariate methods, such as latent factor models, including genomic structural equation modeling (SEM). Genomic SEM tests the fit of *a priori*

defined latent factors to the space of multiple phenotypes[10]. Genomic SEM has been applied successfully to multiple psychiatric[11,12] and cognitive traits[13], and boosts the discovery power of GWAS compared to standard GWAS, while also providing new insights into the overlap of genetic effects across multiple related traits[14]. However, the need to predefine latent factor structures, combined with computational demands of genomic SEM, generally limit the model order to up to ~20 traits. Another method, Multivariate Omnibus Statistical Test (MOSTest) integrates multiple phenotypes in a single powerful GWAS by combining test-statistics in a way akin to meta-analysis[15]. MOSTest derives a single genome-wide signature reflecting the combined effects of each SNP across multiple traits, and it does not look "beyond" this main combined effect. Compared to latent genomic factor models, MOSTest does not inform on which genetic factors modulate different traits in concert. Recent work has explored the application of principal component analysis (PCA) to genome-wide variant z-scores among 220 phenotypes and decomposed a number of latent genomic factors[16]. PCA finds consecutive orthogonal components that explain maximum variance in high-dimensional data. In that sense, given a set of genetically correlated traits, PCA will explain most of the shared variation in as few components as possible. Considering that the genome-wide association profiles of many brain traits are at least partially overlapping, methods like PCA should leverage this share of signal across the traits to derive multivariate genetic components that capture a lot of the signal very efficiently.

Both PCA and ICA are data reduction methods that reduce large matrices of observations to manageable dimensions, through the identification of latent components that capture information that is shared across the observations[17]. While PCA aims to explain the maximum amount of variance using the fewest number of components, ICA is generally more suitable to improve interpretability, and separating distinct sources of signal from noise[18]. In contrast to PCA, where the components and their corresponding weights are orthogonal, ICA components are optimized to be maximally independent but the procedure allows their weights to be non-orthogonal if necessary. By design, ICA seeks to 'unmix' a complex signal, i.e., a signal that is a sum of multiple underlying generative processes, into its sources, which thus become more clearly visible and interpretable[19]. For example, in our GWAS application, the genomic sources of variation in brain structure and function may reflect biological pathways such as myelination, synaptogenesis, cell division, cell death, and neuroinflammation processes that influence various neuroimaging readouts but act relatively independently. Our genomic ICA application is thus based on the premise that those SNPs that influence genes in a coherent biological process impose more similar (hence mutually dependent) impacts on brain traits. In contrast, SNPs in distant biological pathways may have more orthogonal effects since these biological pathways function more independently of each other. This type of latent structure has been previously explored in gene expression data[20,21], but not in the space of genome-wide allelic effects on polygenic traits. Considering the efficiency of ICA in separating signal from noise, some sources may capture systematic 'noise' effects that were incompletely accounted for in the original GWAS, e.g., population stratification or assortative mating. Other sources of variation in the GWAS output may be less specific at a molecular level, but more environmentally mediated. For example, sociodemographic factors or lifestyle factors that are themselves also heritable[12,22].

In our previous, preliminary application of genomic ICA we provide evidence suggesting that many of the independent components indeed mapped to one or more of such underlying biological pathways[7]. In particular, we extracted independent components from GWAS output of thousands of brain-related traits from UK Biobank[17]. The independent genomic components we thus identified were driven by

different sets of largely modality-specific neuroimaging features and showed enhanced sensitivity to biological pathways compared to the original GWAS output. In this proof-of-concept work, we thus demonstrated the potential utility of genomic ICA to facilitate interpretation of multifactorial GWAS signals in terms of biological mechanisms. This in turn may have future applications for stratification of heterogeneous patient populations.

Here, we apply genomic ICA to the more recent UK biobank GWAS release of 2,240 brain MRI traits, including larger, non-overlapping discovery (n=22,138) and replication (n=11,086) samples[23] to test the stability and reliability of genomic source decomposition. We further extend genomic ICA and compare it with genomic principal component analyses (PCA). We assess the reproducibility of the independent and principal genomic components across split-half subsets of traits within a single sample's GWAS output (intra-sample reproducibility), and across the GWAS outputs of the two independent samples. Finally, we provide a head-to-head comparison of genomic PCA and genomic ICA reproducibility with the reproducibility of the original univariate GWAS.

Methods:

Data

Genome-wide association study (GWAS) summary statistics were acquired from the Oxford Brain Imaging Genetics (BIG-40) database. This open-source repository contains the results of ~4,000 GWASs performed using brain magnetic resonance imaging (MRI-) derived phenotypes (IDPs) in UK biobank[24]. For the current study, we considered summary statistics of two independent rounds of GWAS in non-overlapping discovery and replication samples, encompassing 22,138 and 11,086 subjects with an identical set of brain IDPs performed by Smith et al.[23] We excluded node-to-node functional connectivity IDPs due to their low heritability[25], leaving a total of 2,240 IDPs for further analysis. Notably, the amplitudes of functional signal fluctuations and six ICA-derived 'global' measures of functional connectivity were included, as they have been shown to be heritable features[23,25]. Other IDP classes included quantitative parameters extracted from T1-weighted MRI, diffusion MRI, susceptibility-weighted imaging (SWI), fluid-attenuated inversion recovery (FLAIR), task fMRI and quality-control procedures. For further details on the imaging genetics pipelines of UK biobank please refer to the main publications[23,25] (https://open.win.ox.ac.uk/ukbiobank/big40).

Clumping

We applied genome-wide SNP clumping to reduce redundancy and to remove excessive local SNP-dependencies stemming from linkage disequilibrium (LD). The smallest p-value of each SNP across the GWASs of 2,240 brain IDPs was considered for clumping. A genomic window size of one mega bases, a lead variant p-value threshold of 10^{-5} and a linkage disequilibrium (LD) threshold of $r^2>0.3$ were used as clumping parameters. LD was estimated in a random subsample of 10,077 Caucasian UK biobank subjects (data field number 22006). This procedure reduced the total number of genome-wide SNPs from 17,103,079 to 157,893, effectively minimizing the impact of LD on local SNP-to-SNP correlations while keeping brain-related lead variants within each LD block for downstream analysis.

Principal and Independent component analysis

The vectors of genome-wide SNP effect sizes (i.e., the GWAS regression parameter estimates) were concatenated across all IDPs, generating an m×n brain-wide genome-wide matrix of imaging IDPs (m=2,240) and genetic variants (n=157,893 clumped SNPs). We decomposed this brain-wide genome-

wide matrix of SNP effect sizes by MELODIC, which is a probabilistic ICA algorithm maximizing non-Gaussianity in the reconstructed sources[26]. In the standard pipeline, MELODIC starts by applying principal component analysis (PCA) to data, and rotates these predefined number of components to optimize a measure of independence.

The m×n brain-wide genome-wide matrix was decomposed by MELODIC along the IDP dimension into a maximum of 50 independent components of SNP effect sizes by PCA. This dimension was chosen by a visual assessment of the scree plots, indicating that dimension 50 provides a good balance between explained variance and model complexity (Figure 1, 60.2% SNP effect variance explained). In addition to PCA, MELODIC ICA was applied at multiple dimensions, including 5, 10, 25, and 50 in the discovery and replication samples to extract genomic components, hereafter referred to as independent genomic sources. The global signal removal option of MELODIC was disabled (using the --keep_meanvol flag) since SNP effect size distribution is already centered on zero under the null assumption of a lack of biological association between SNPs and traits. SNP variance normalization across IDPs was disabled (--vn) to preserve the magnitude of allelic effects in the genomic components.

Intra-sample reliability

We implemented a split-half cross-validation scheme to study the robustness of ICA against the inclusion or exclusion of brain IDPs in genomic source decomposition. Focusing on the larger discovery sample (n=22,138), we randomly shuffled the rows of the IDP×SNP matrix and then split this matrix along the IDP rows into two halves. PCA and ICA were then applied to each of the split halves (i.e., m=1,120 IDPs) using identical parameters described above. Two types of measures were then used to assess the reproducibility of genomic sources across all component pairs of the two split-halves; (1) to assess the sparse part of the genomic sources, i.e., those variants that strongly contribute to each component's multivariate effect compared to the background 'noise'[26], we used overlap coefficient in the component tails, i.e., SNPs with PCA or ICA component weights of $|z|>2$, similar to a previous work[27]; and (2) to investigate the non-sparse, global genomic signals of all variants contributing to each component, Pearson's correlation coefficients of SNP-wise weights were used without component thresholding. Matrix re-shuffling and split-half PCA and ICA was repeated k=20 times to obtain an empirical distribution of the above two reliability metrics, generating a total of 3,600 ICA and 2,000 PCA components across all dimensions and k-folds. Across the folds, reliability of each component was assessed by choosing a 'matching' component in the second half of the dataset, which was determined based on maximum pairwise correlation or overlap coefficient across the source decompositions.

Cross-sample reliability of the genomic sources

To investigate the reliability of genomic sources across independent GWAS samples, the genomic sources were compared across the discovery and replication samples using the same reproducibility metrics described above. Both full-matrix (non-split) PCA and ICA components and their split-half versions were included to obtain a distribution of the reproducibility metrics across the k-folds.

Reproducibility of the univariate SNP effects

The reproducibility of SNP effect sizes in the original (i.e. standard univariate) GWASs was assessed across the discovery and replication samples using both Pearson's correlation (r_{SNP}) of variant effect sizes and tail overlap coefficient ($|z|>2$). Maximum r^2_{SNP} across the samples is theoretically determined by additive SNP heritability (h^2_{SNP}) of each trait, and this measure provides a benchmark for assessing the reproducibility of PCA and ICA genomic components.

Results

Reproducibility of the raw univariate SNP effects

The reproducibility of 'raw' SNP effect sizes from univariate GWASs are shown in Figure 1. The average reproducibility in univariate GWASs was 0.09±0.05 (Pearson's r; mean ± standard deviation) across the discovery and replication samples. The highest GWAS SNP effect reproducibility was observed for global structural MRI features, including *total grey-matter volume* as the top-ranking feature among 2,240 IDPs (r=0.25), followed by *total cortical volume* in hemispheres (r_{right}=0.24 and r_{left}=0.23), *average cortical thickness* (r_{both}=0.23) and *total cortical surface area* (r_{both}=0.22). Diffusion MRI measures in large tracts showed high genetic reproducibility as well, spanning *intracellular volume fraction (ICVF)* in the *inferior longitudinal fasciculus* and *inferior fronto-occipital fasciculus* (r_{right}=0.23 and r_{left}=0.24) and *mean diffusivity (MD)* in *anterior corona radiata* (r_{both}=0.23). IDPs extracted from other modalities showed weaker SNP effect reproducibility, probably owing to lower SNP heritability and/or higher polygenicity of these traits (maximum r: 0.15 for T1 intensity-derived features, maximum r: 0.04 for SWI features, and maximum r: 0.04 for the six ICA-derived functional connectivity meta-IDPs).

The highest overlap coefficient in SNP effect tails (|z|>2) were mostly observed in diffusion MRI-derived features. Notably, of the first 100 IDPs with the highest tail overlaps, 95 were diffusion MRI IDPs. The top-ranking ones included *ICVF* in the *inferior fronto-occipital fasciculi* (O_{right}=0.27 and O_{left}=0.28) and *inferior longitudinal fasciculi* (O_{both}=0.27). The five non-diffusion IDPs among the top 100 were *global mean cortical thickness* (O_{both}=0.27), *global volume and intensity of the 5th ventricle* (O=0.26) and *global intensity of non-white-matter hypointensities* (O=0.26).

Dimensionality and explained variance of PCA-derived genomic components

The first five principal components extracted from the GWAS of 2,240 IDPs captured 32.0% of the variance in SNP effect sizes. Increasing the number of PCs to 200 resulted in 79.1% of the variance explained (Figure 2). The first ten PCs were strongly reproducible within samples in 20-fold split-halves (Pearson's r=0.82±0.15, range: 0.46-0.99) but their cross-sample reproducibility dropped to r=0.16±0.07 (range=0.06-0.31; Figure 3). Notably, this was nevertheless a major improvement over the cross-sample reproducibility of univariate GWASs across samples (r=0.09±0.05; see the previous section). For example, the first genomic principal component was more reproducible than the GWAS of *total grey-matter volume*, which was the most reproducible univariate GWAS across all IDPs (r_{PCA}=0.30 vs. r_{IDP}=0.25). The first genomic PC showed a relatively strong tail overlap across samples (O_{PCA}=0.27) compared to the tail overlap of univariate GWASs (O_{IDP}=0.23±0.01). The subsequent principal components (#11-50), which explain successively less variance were less reproducible both within (r=0.41±0.14; range: 0.16-0.89) and across samples (r=0.04±0.02; range: 0.01-0.09; Figure 3).

Reproducibility of the ICA-derived genomic sources

The stability and reproducibility of genomic ICA analyses are shown in Figures 4 and 5. ICA-decomposed genomic components showed higher within-sample reproducibility at lower ICA dimensions (r_{ICA-5}=0.81±0.16, r_{ICA-10}=0.74±0.15, r_{ICA-25}=0.57±0.15, r_{ICA-50}=0.41±0.13). The same trend was observed for cross-sample reproducibility of the genomic components (r_{ICA-5}=0.20±0.3, r_{ICA-10}=0.14±0.04, r_{ICA-25}=0.08±0.03, r_{ICA-50}=0.05±0.02). In general, PCA tended to capture most of the variance by the first 1-3 components but showed a substantial drop in the reproducibility of the lower-variance components,

whereas ICA, especially at lower dimensions, applied rotations to the data in a way that resulted in multiple components to remain reproducible.

Discussion

Each GWAS opens a new window into candidate genes and biological mechanisms that shape a trait. The sheer complexity of the web of mechanisms underlying polygenic inheritance limit the immediate mechanistic interpretation of GWAS, for example towards translational research and drug targeting. Our genomic ICA and PCA methods take advantage of two properties of brain MRI traits that are generally considered a challenge: polygenicity and pleiotropy. We use these multivariate algorithms to reduce the observed GWAS results of more than 2,000 brain traits into a small set of multivariate genomic components. Specifically, we applied PCA and ICA to abstract the high-dimensional brain-wide genome-wide signal into latent factors in a data-driven way, i.e., without relying on *a priori* curated pathways or gene-set annotations. We identified latent structure in the matrix of genome-wide effects that efficiently explained the pleiotropy in the observed GWAS patterns. Within samples, both PCA and ICA components were highly reproducible across sets of traits. The reproducibility across independent datasets was moderate, but significant, for both genomic PCA and ICA and easily surpassed that of the raw univariate GWAS output. Notably, the most reproducible components were those of the first ~3 genomic principal components (PCs) that explain the highest amount of variance (i.e., *salient features*), while the reproducibility of genomic ICA was less good, especially for higher dimensional decompositions. Theoretically, it can be predicted that the reproducibility of the (lower-order) independent components – that capture less variance and are typically more locus-specific – will increase as GWAS discovery sample sizes continue to increase.

While optimizing reproducibility is favorable, it has not been the only goal of our genomic source decomposition. Rather, we aimed at obtaining a latent representation of the genome-wide variant effect sizes that aligns more closely to coherent biological processes[7]. Our comparison of PCA and ICA demonstrated that with the exception of the first three (salient) PCA components, the rest of the PCA and ICA components are comparable with regards to cross-sample reproducibility. PCA tended to capture most of the variance by the first 1-3 components but showed a substantial drop in the reproducibility of the lower-variance components, whereas ICA, especially at lower dimensions, applied rotations to the data that resulted in multiple components to remain relatively reproducible across samples, probably as a result of distributing biological signal across multiple independent sources. Notably, genomic ICA was more reproducible at the lowest dimension five with a cross-sample reproducibility of r~0.2, which is comparable to the reproducibility of PCA components and both were better than the raw univariate GWAS effect sizes. Together, these results are in line with the theoretical assumption that unlike PCA, ICA is aimed to recover the original processes (a.k.a. hidden 'sources') that generate a mixed signal, while principal components recover orthogonal axes of maximum variance. Because independence is a stricter criterion than uncorrelatedness, principal components can comprise a combination of co-dependent patterns within a single component, which in ICA would be decomposed into multiple sources. Due to this lack of *source unmixing* in PCA, the principal components we obtained here probably represent a convoluted mixture of many biological processes that affect the brain.

The top reproducible IDPs in univariate GWAS were the more 'global' measures, such as total cortical volume and total cortical surface area. We speculate that these global features tend to be more reproducible across the samples for two potential reasons; global imaging features summarize a large

collection of brain subregions, resulting in signal 'averaging', relatively less noise and higher phenotypic reproducibility. In addition, the genetic profile of a global feature is a combination of many subregions, each one with a distinct genetic profile. Therefore, the genome-wide signal of global MRI traits may be more diverse and less sparse than that of subregions. Indeed, higher heritability have been reported previously for global diffusion MRI features[28]. ICA may be theoretically sound in separating hidden sources in such a convoluted genome-wide signal, as it reconstructs a number of maximally independent genomic components from a mixed signal. With further increases in discovery GWAS sample sizes and the ensuing increased precision in the GWAS observations, we expect independent genomic components to be able to separate independent and coherent biological processes (e.g. at the cellular scale) and from sources of structured noise[29] such as population stratification effects more and more reliably.

The flexibility of PCA and ICA allow applications to different modalities[33,34], bringing new opportunities to other non-GWAS data explorations, such as transcriptome and epigenome data. Another promising path we are exploring is the projection of the genomic components into individual subjects' genotypes and using the subject-wise component-scores in building individual genomic component scores, similar to polygenic risk scores, to stratify heterogeneous populations, for prediction models, towards personalized medicine. As GWAS sample sizes of MRI measures and other traits continue to grow, these multivariate components will likely become more reliable and interpretable, warranting future investigations into their alignment to cell types, neurodevelopmental processes and brain homeostasis pathways.

Limitations

Genomic ICA can better capture stable and biologically relevant components when more *views* of the underlying sources are available in the form of GWAS results across multiple traits. Although we used various brain MRI modalities for genomic source decomposition, adding other non-MRI brain-related phenotypes may help to recover more genomic sources with translational potential, such as measures related to disease biomarkers, behavioral traits and neurological and psychiatric conditions. The strong phenotypic correlations between many MRI traits may be more suitable to a lower-dimensional decomposition that we initially aimed to achieve. Future works may need to delve deeper into optimizing genomic source decomposition e.g. using modality-aware data fusion by ICA[34]. Another limitation of our approach is our restriction to a European ancestry sample, since this is what the original discovery GWAS had available. Some of the multivariate components describing the GWAS-brain associations may be more or less sensitive to specific genetic ancestries, or socio-cultural contexts. We eagerly await more diversity in the GWAS discovery data that also includes MRI traits, but it is appreciated that the data acquisition costs are high and not all standard methods are (yet) applicable to multiple ancestries. Even for more standard methods and more affordable phenotyping, interracial and intercultural diversity of GWAS data and (especially) results remains limited.

Conclusion

We derived genomic components describing the pleiotropic and polygenic influences on 2,240 brain MRI traits. Genomic PCA identified highly reproducible genomic effects across many brain traits in up to three factors. Genomic ICA obtained comparable reproducibility at lower dimensions and captured mutually independent effects that may highlight distinct biological pathways. Genomic ICA and PCA can help improving the signal-to-noise and interpretability of large GWAS catalogues across biobanks by

leveraging the cross-trait similarity of genome-wide associations and separating structured noise from meaningful biological signal.

References

1 Pizzagalli, F. *et al.* The reliability and heritability of cortical folds and their genetic correlations across hemispheres. *Communications Biology* **3**, 510, doi:10.1038/s42003-020-01163-1 (2020).

2 Eyler, L. T. *et al.* Genetic and environmental contributions to regional cortical surface area in humans: a magnetic resonance imaging twin study. **21**, 2313-2321 (2011).

3 Hibar, D. P. *et al.* Novel genetic loci associated with hippocampal volume. **8**, 1-12 (2017).

4 Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* **24**, 737-745, doi:10.1038/s41593-021-00826-4 (2021).

5 Wendt, F. R., Pathak, G. A., Tylee, D. S., Goswami, A. & Polimanti, R. J. C. S. Heterogeneity and Polygenicity in Psychiatric Disorders: A Genome-Wide Perspective. **4**, 2470547020924844 (2020).

6 Matoba, N., Love, M. I. & Stein, J. L. J. H. B. M. Evaluating brain structure traits as endophenotypes using polygenicity and discoverability. (2020).

7 Soheili-Nezhad, S., Beckmann, C. F. & Sprooten, E. J. b. Independent Genomic Sources of Brain Structure and Function. (2021).

8 Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications* **11**, 5900, doi:10.1038/s41467-020-19653-5 (2020).

9 Hemani, G. *et al.* Phantom epistasis between unlinked loci. *Nature* **596**, E1-E3, doi:10.1038/s41586-021-03765-z (2021).

10 Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour* **3**, 513-525, doi:10.1038/s41562-019-0566-x (2019).

11 Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* **3**, 513-525, doi:10.1038/s41562-019-0566-x (2019).

12 Marees, A. T. *et al.* Genetic correlates of socio-economic status influence the pattern of shared heritability across mental health traits. *Nature Human Behaviour* **5**, 1065-1073, doi:10.1038/s41562-021-01053-4 (2021).

13 Warrier, V. *et al.* Social and non-social autism symptoms and trait domains are genetically dissociable. *Communications Biology* **2**, 328, doi:10.1038/s42003-019-0558-4 (2019).

14 Thorp, J. G. *et al.* Genetic heterogeneity in self-reported depressive symptoms identified through genetic analyses of the PHQ-9. *Psychological Medicine* **50**, 2385-2396, doi:10.1017/S0033291719002526 (2020).

15 van der Meer, D. *et al.* Understanding the genetic determinants of the brain with MOSTest. *Nature Communications* **11**, 3512, doi:10.1038/s41467-020-17368-1 (2020).

16 Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics* **53**, 1415-1424, doi:10.1038/s41588-021-00931-x (2021).

17 Smith, S. M. *et al.* Enhanced Brain Imaging Genetics in UK Biobank. *bioRxiv*, 2020.2007.2027.223545, doi:10.1101/2020.07.27.223545 (2020).

18 Comon, P. J. S. p. Independent component analysis, a new concept? **36**, 287-314 (1994).

19 Hyvärinen, A. Survey on independent component analysis. (1999).

20 Wang, W. *et al.* Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Res* **49**, e54-e54, doi:10.1093/nar/gkab089 %J Nucleic Acids Research (2021).
21 Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T. & Huang, X. A review of independent component analysis application to microarray gene expression data. *Biotechniques* **45**, 501-520, doi:10.2144/000112950 (2008).
22 Abdellaoui, A. *et al.* Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour* **3**, 1332-1342, doi:10.1038/s41562-019-0757-5 (2019).
23 Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* **24**, 737-745, doi:10.1038/s41593-021-00826-4 (2021).
24 Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400-424, doi:https://doi.org/10.1016/j.neuroimage.2017.10.034 (2018).
25 Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210-216, doi:10.1038/s41586-018-0571-7 (2018).
26 Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging* **23**, 137-152 (2004).
27 Pruim, R. H., Mennes, M., Buitelaar, J. K. & Beckmann, C. F. J. N. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. **112**, 278-287 (2015).
28 Kochunov, P. *et al.* Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data. **111**, 300-311 (2015).
29 Pruim, R. H. *et al.* ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. **112**, 267-277 (2015).
30 Hyvarinen, A. J. I. t. o. N. N. Fast and robust fixed-point algorithms for independent component analysis. **10**, 626-634 (1999).
31 Le, Q., Karpenko, A., Ngiam, J. & Ng, A. J. A. i. n. i. p. s. ICA with reconstruction cost for efficient overcomplete feature learning. **24**, 1017-1025 (2011).
32 Torrey, L. & Shavlik, J. in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* 242-264 (IGI global, 2010).
33 Gong, W., Bai, S., Zheng, Y.-Q., Smith, S. & Beckmann, C. J. b. Supervised Phenotype Discovery from Multimodal Brain Imaging. (2021).
34 Groves, A. R., Beckmann, C. F., Smith, S. M. & Woolrich, M. W. J. N. Linked independent component analysis for multimodal data fusion. **54**, 2198-2217 (2011).

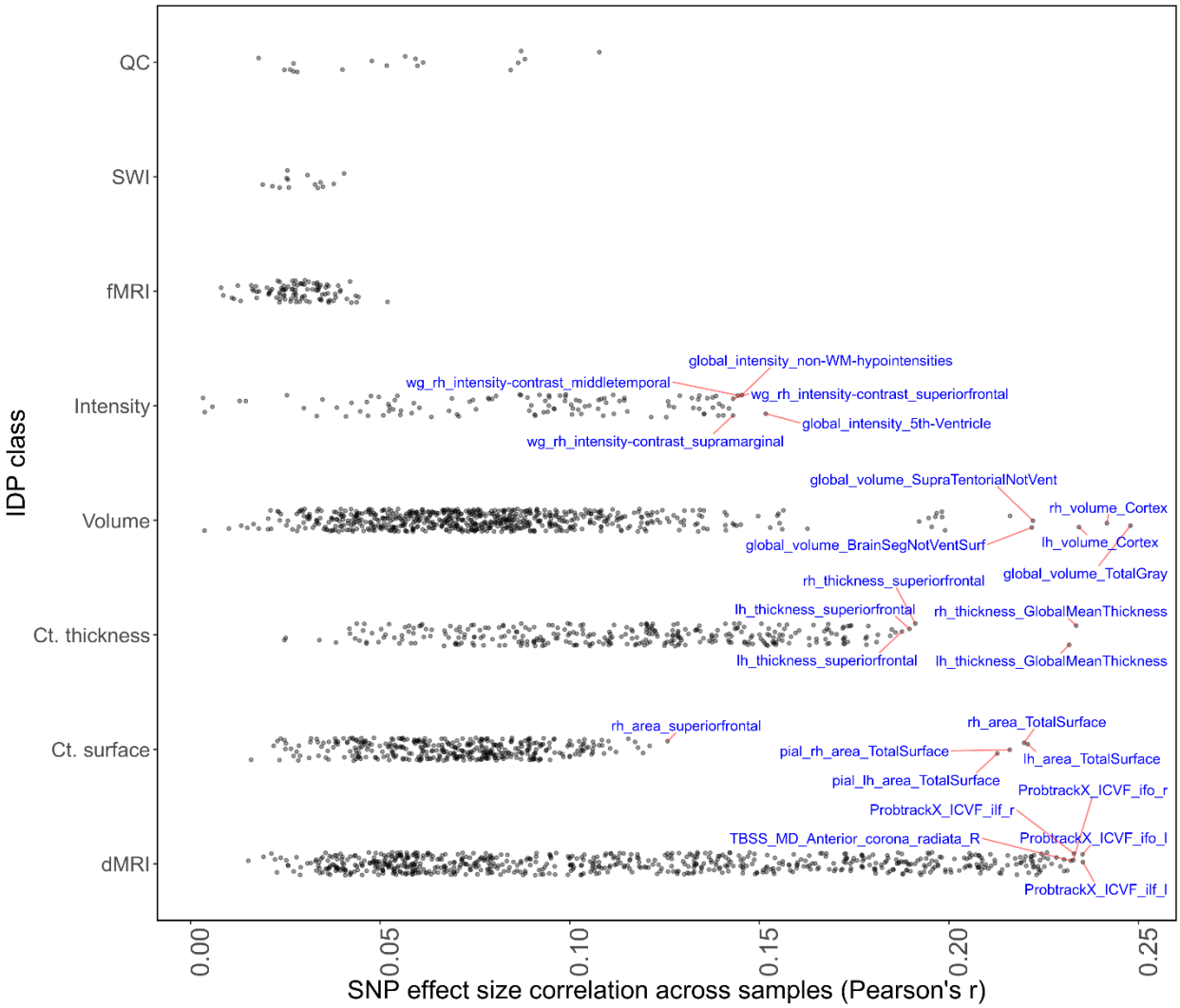


Figure 1. Reproducibility of the 'raw' GWAS SNP effect sizes (n=157,893 clumped variants) across the two samples.

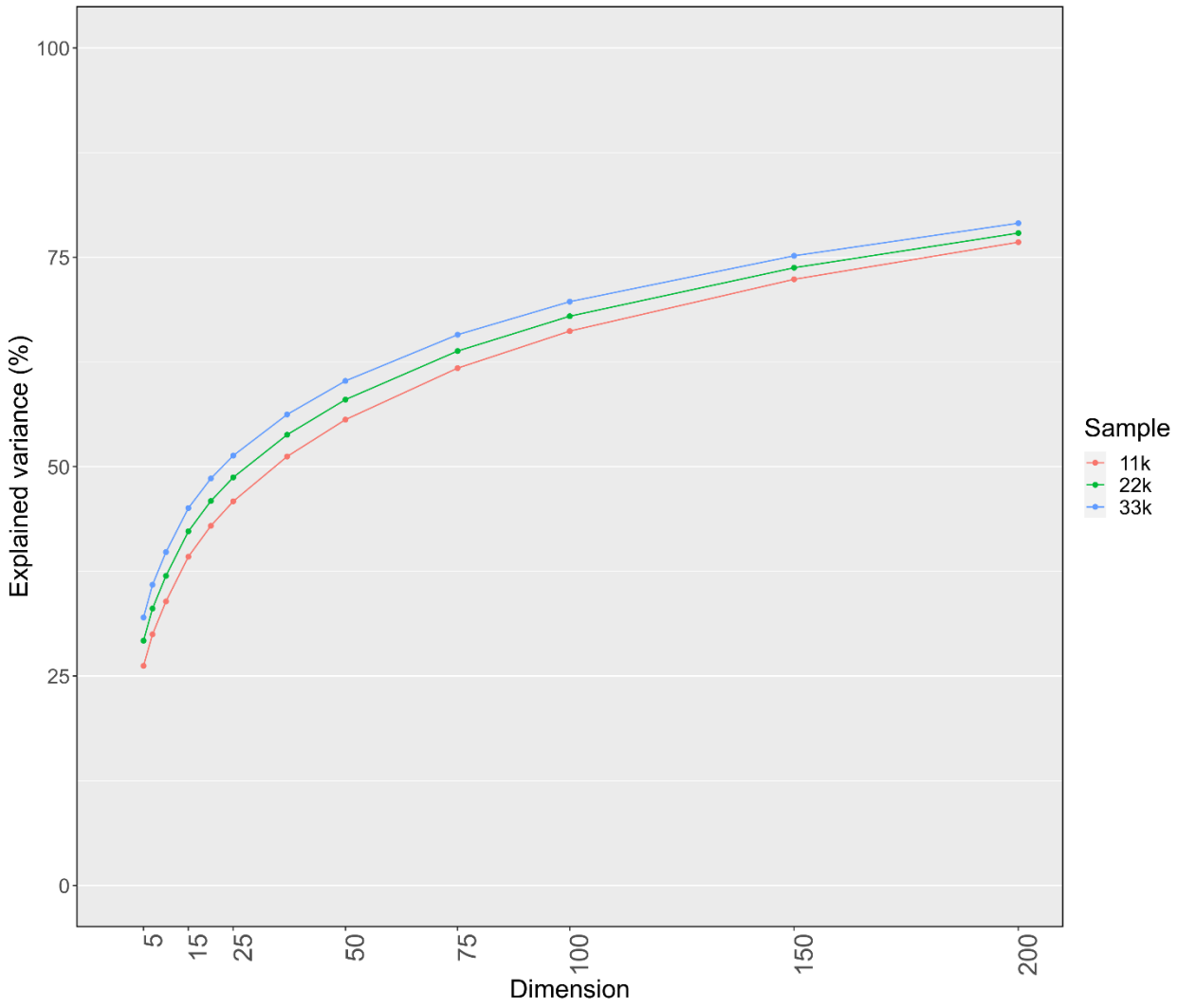


Figure 2. Variance explained by genomic components at ICA dimensions

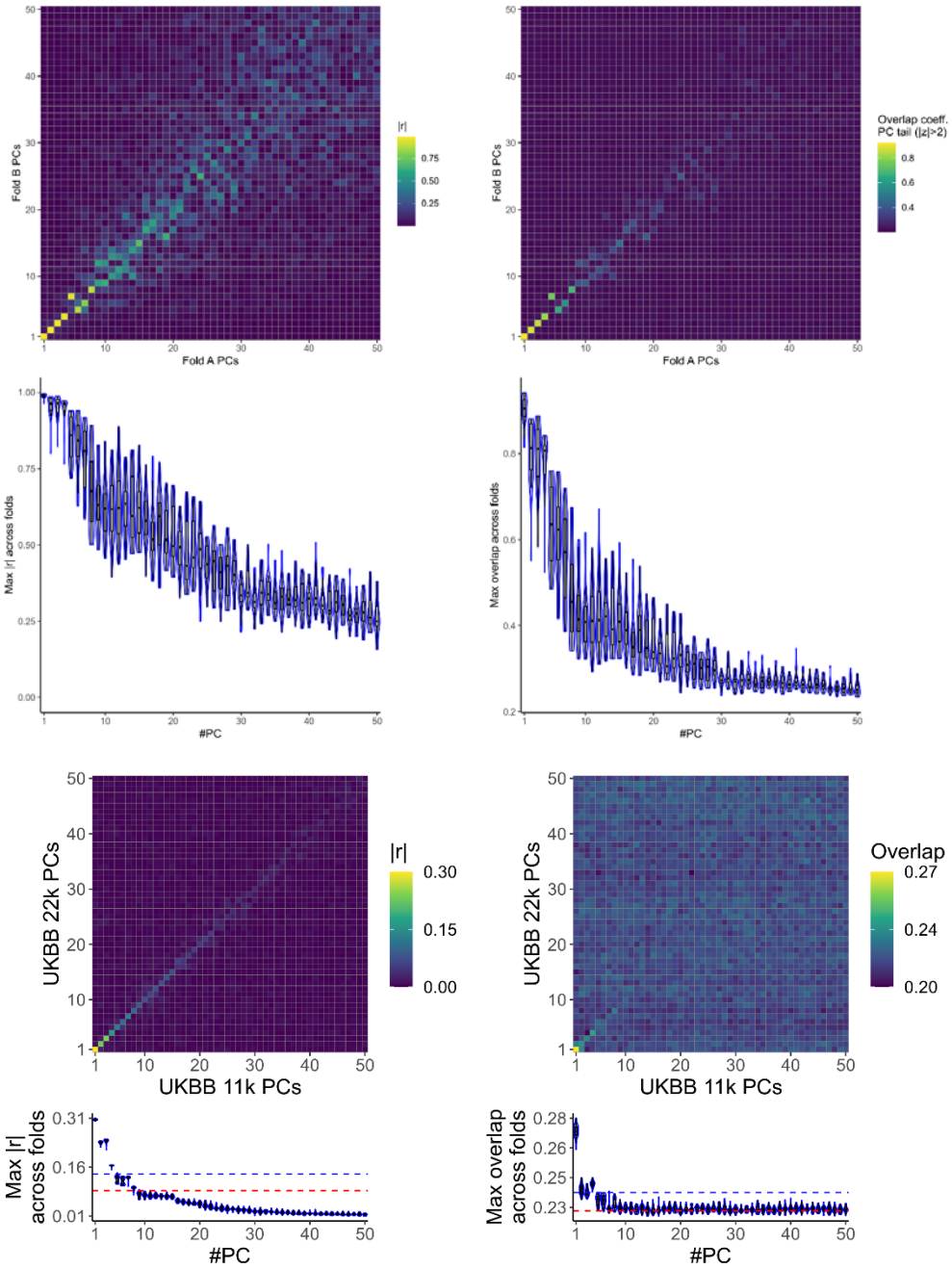


Figure 3. Intra- (top) and inter-sample (bottom) reproducibility of PCA derived genomic components (split-half; k=20 folds; left panels: correlation coefficient; right panels: overlap coefficient; dashed blue lines: maximum reproducibility of raw univariate SNP effects; dashed red lines: average reproducibility of raw univariate SNP effects)

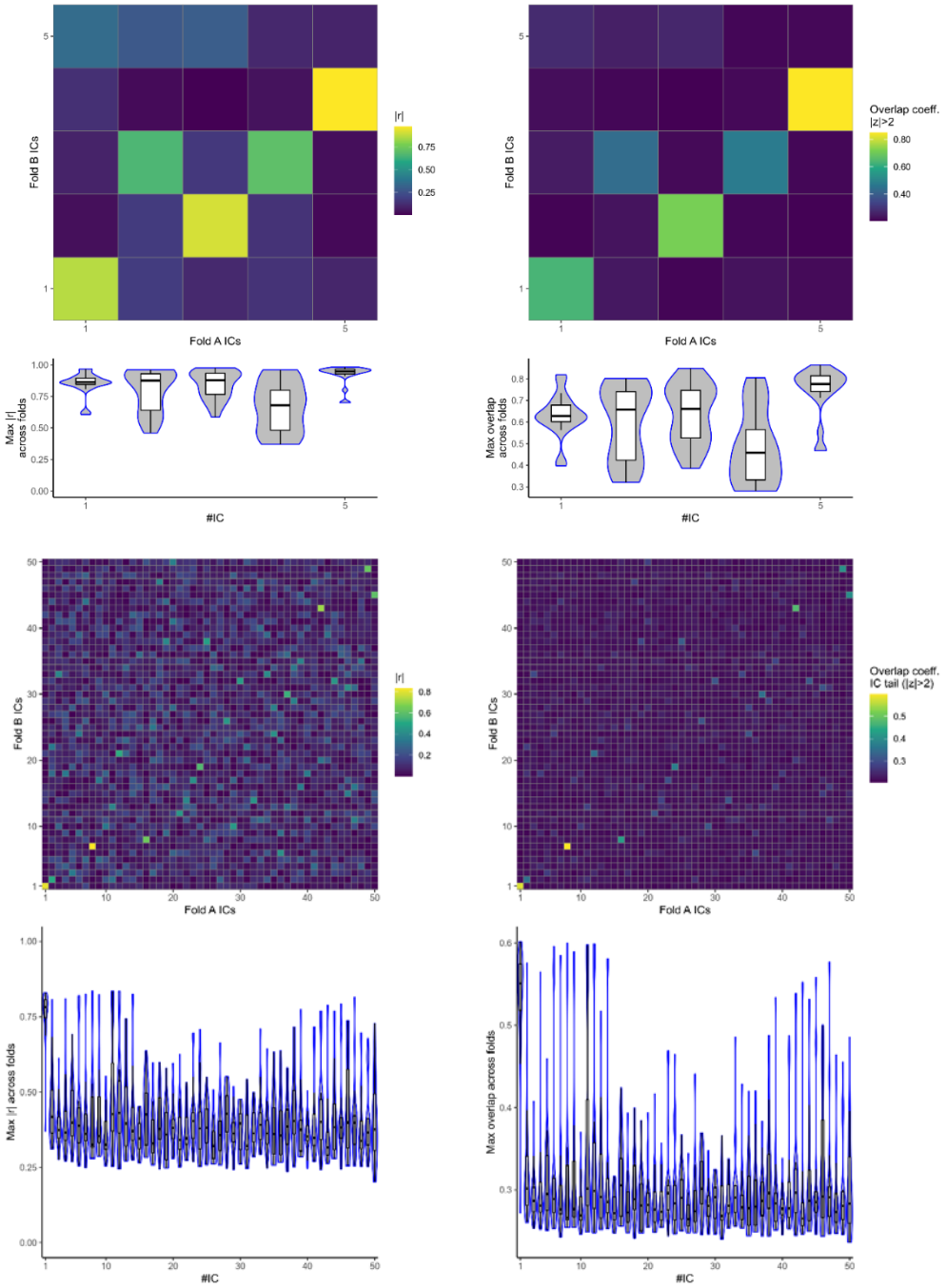


Figure 4. Intra-sample reproducibility of genomic sources. ICA dimensions 5 and 50; split-half; k=20 folds; left panels: correlation coefficient; right panels: overlap coefficient (for other dimensions refer to supplementary figure 1).

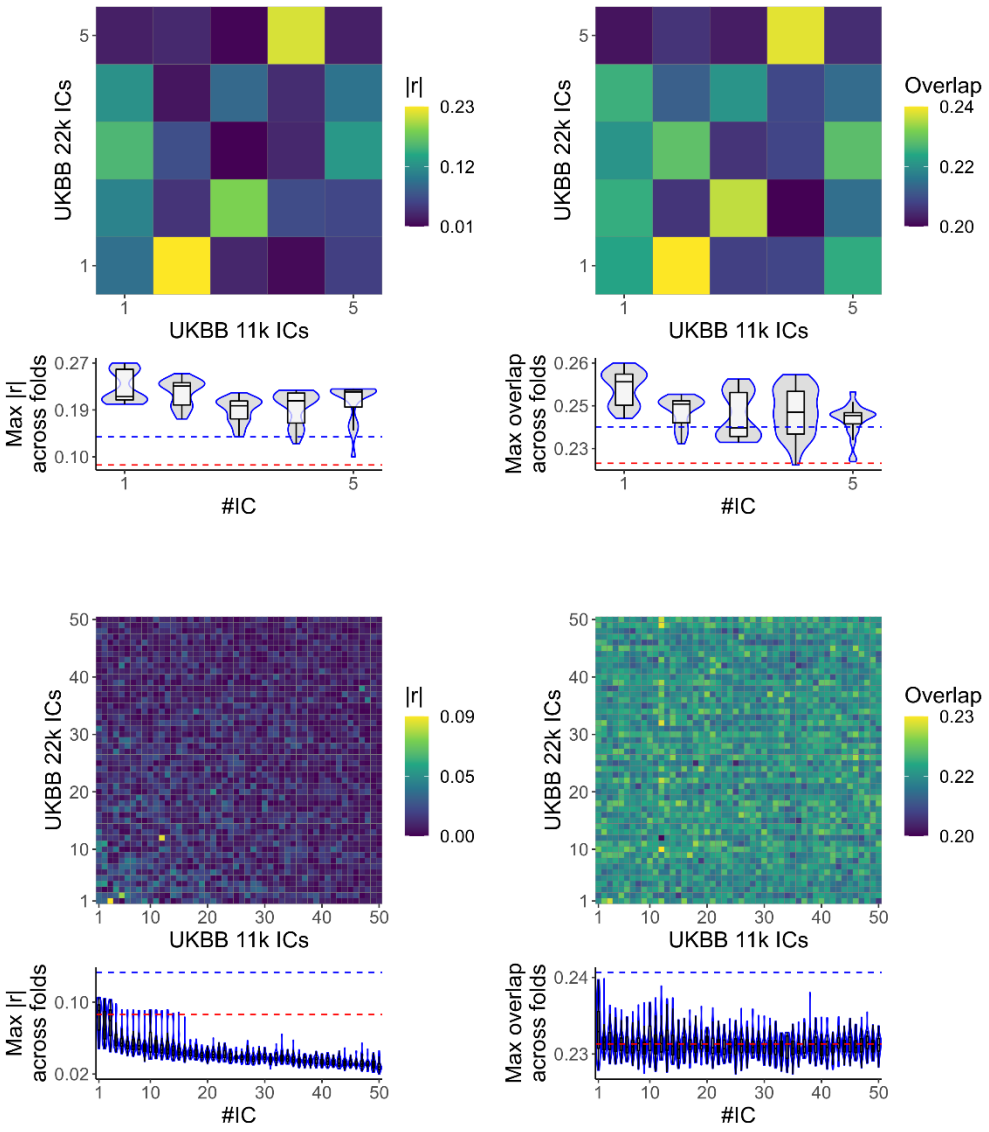
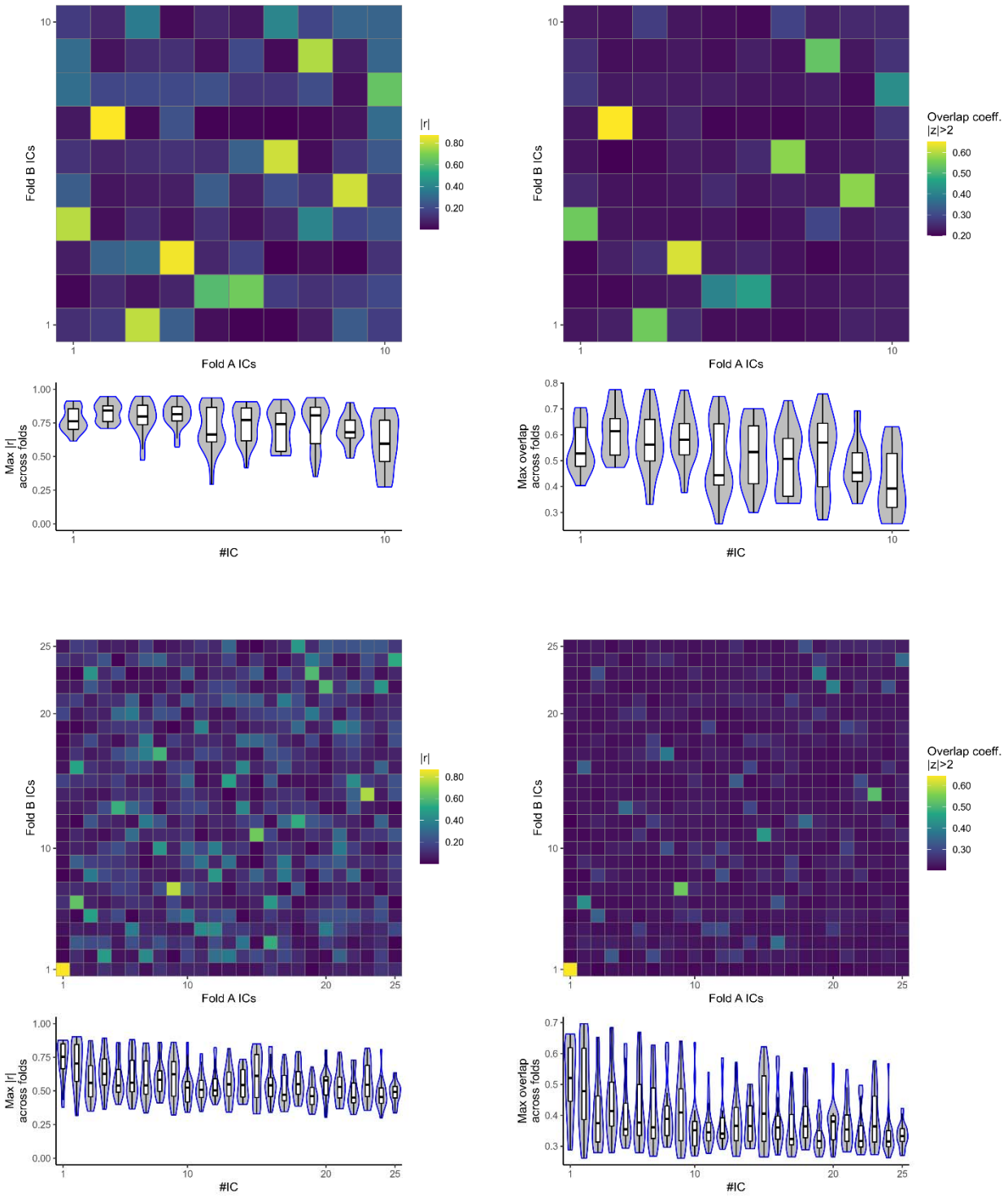
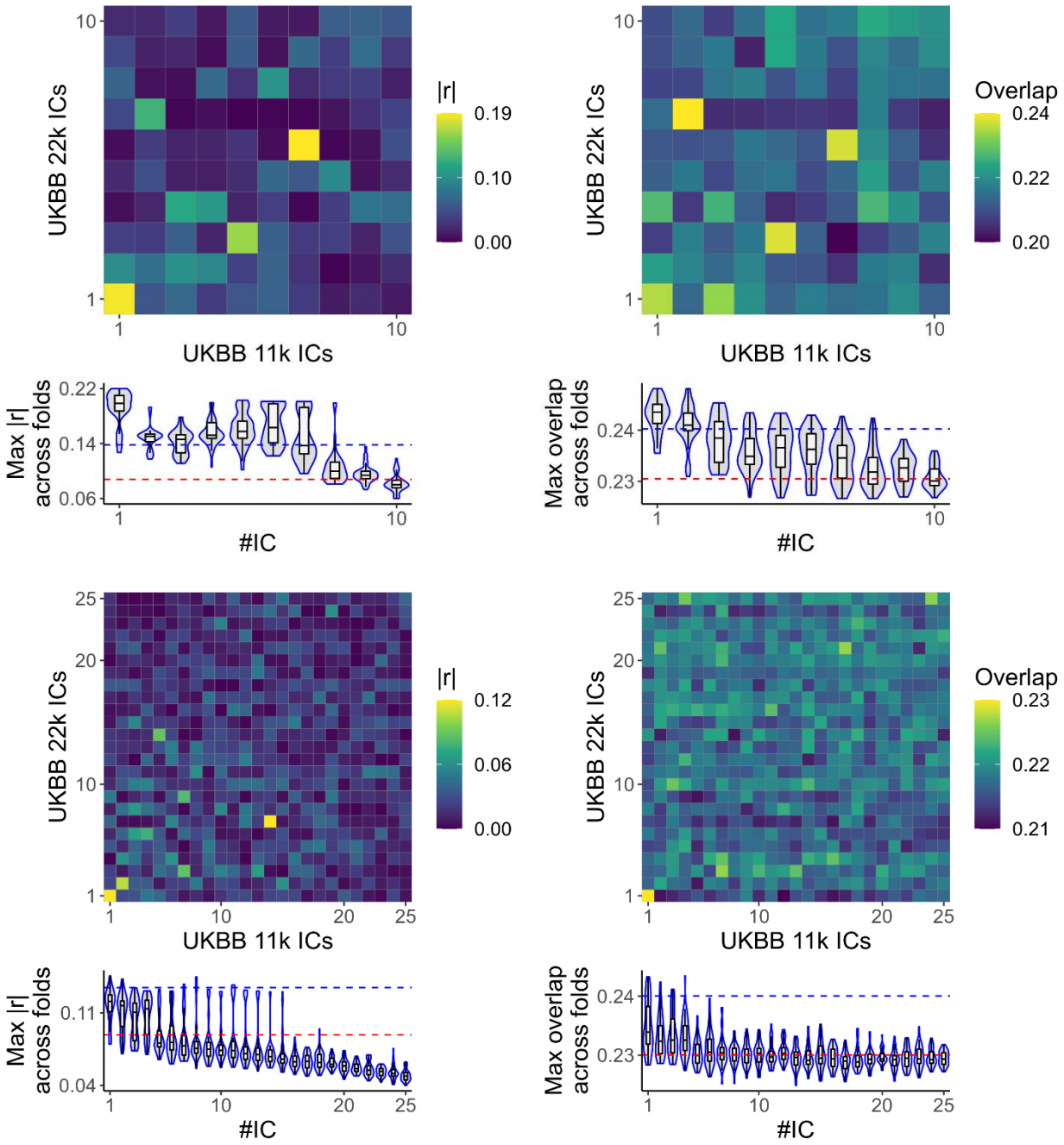


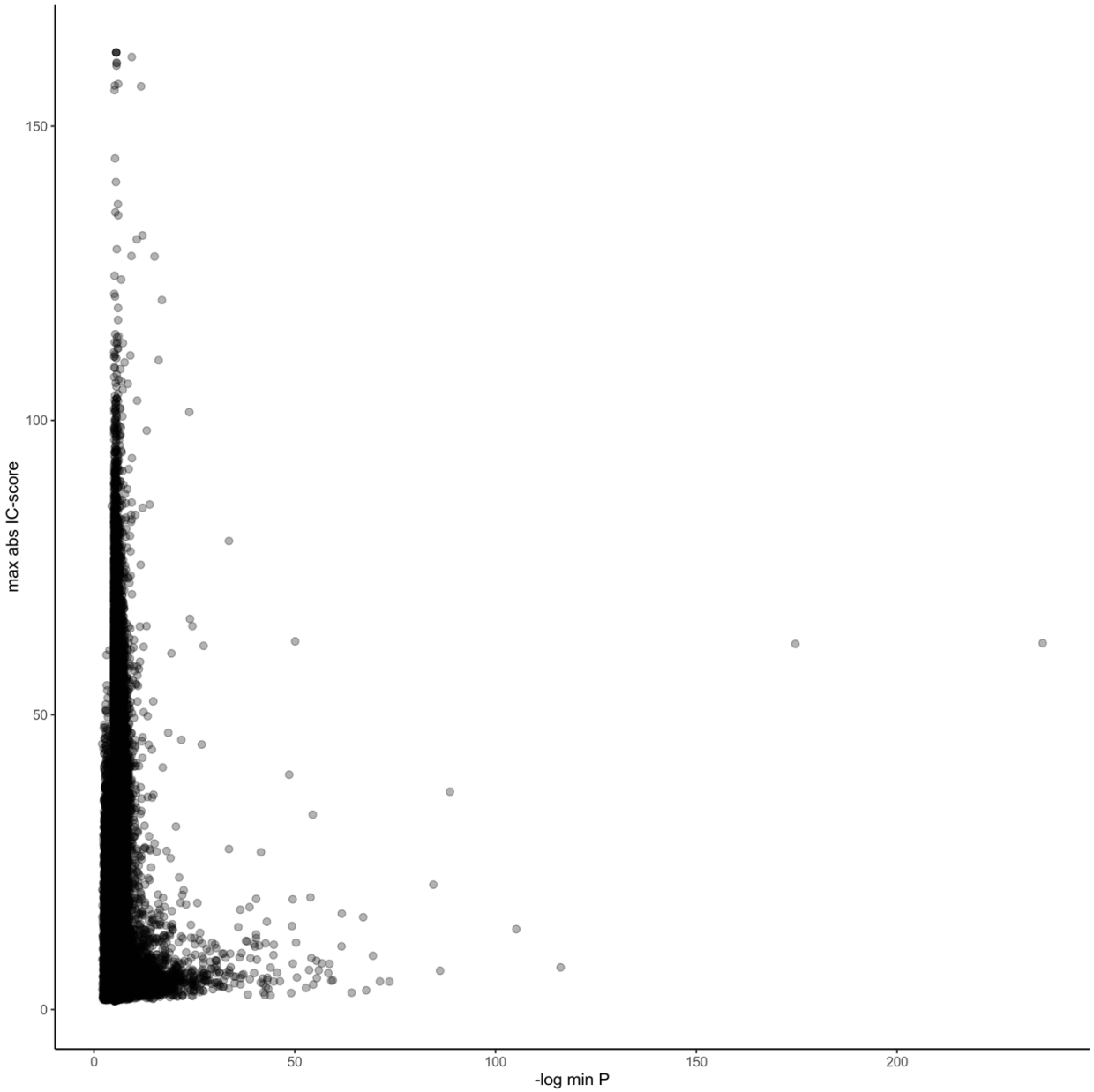
Figure 5. Inter-sample reproducibility of genomic sources. ICA dimensions 5 and 50; split-half; k=20 folds; left panels: correlation coefficient; right panels: overlap coefficient; dashed blue lines: maximum reproducibility of raw univariate SNP effects; dashed red lines: average reproducibility of raw univariate SNP effects (for other dimensions refer to supplementary figure 2).



Supplementary figure 1. Intra-sample reproducibility of genomic sources. ICA dimensions 10 and 25; split-half; k=20 folds; left panels: correlation coefficient; right panels: overlap coefficient



Supplementary figure 2. Inter-sample reproducibility of genomic sources, ICA dimensions 10 and 25; split-half; k=20 folds; left panels: correlation coefficient; right panels: overlap coefficient; dashed blue lines: maximum reproducibility of raw univariate SNP effects; dashed red lines: average reproducibility of raw univariate SNP effects



Supplementary figure 3. minimum SNP GWAS p-values across 2,240 IDPs and ICA scores. Variants with higher ICA scores (normalized to unit standard deviation) are not necessarily those with smaller GWAS p-values, indicating that the multivariate genomic effects are not captured by the standard univariate GWAS.